CHAPTER III

METHODOLOGY

Sample

Students enrolled in the physics class of the Degree Foundation Studies in Science and Engineering in a private college in Klang, Selangor served as sample for the study. These students, who have passed the *S.P.M* examination, consist of 18 males and 15 females with ages between 18 and 19 years. They have varied background and come from different states in Malaysia. Ninety-four percent of the students are Malaysians of Chinese origin while six percent are of Indian origin.

At the beginning of this study, these students have already completed the first semester of their physics course covering 18 weeks of lectures. They have also sat for the semester examination which is a summative test consisting of test items sampled from the syllabus content for that semester. These include General Physics, Mechanics, Electricity and Magnetism.

After a one week break, these students continue with the second semester of the physics course which consist of 19 weeks of instruction in Waves, Quantum Physics, Atomic Physics and Nuclear Physics. The aggregate of all the test scores for the semester contributes 30 percent to the overall semester grade. Therefore the students are extrinsically motivated to sit for these tests. On completion of this semester, the students will sit for the final semester examination.

Procedure

The procedures used for the collection of data to answer each of the six research questions are as follow:

(1) What are the effects of an instructional strategy based on diagnostic assessment on physics achievement scores in the classroom?

This research question is answered after a longitudinal study by the researcher. This involves applying the diagnostic assessment model to the experimental group of students for the duration of the 19-week semester. The study involves two phases:

(1) Construction phase, and

(2) Application phase.

Construction Phase

This phase is the preparation phase before the researcher applies the diagnostic assessment strategy in the classroom at the commencement of the second semester of the course. This involves organising the physics course curriculum into modules. Each model was then organised into units of instruction. A set of instructional behavioural objectives or intended learning outcomes is constructed for each of the units (Appendix F). These instructional objectives are stated using the classifications of Bloom's taxonomy of educational objectives in the cognitive domain. They include both the lower cognitive skills of knowledge and comprehension, and the higher levels of application, analysis, synthesis and evaluation.

The researcher constructed criterion-referenced tests for each of the units of instruction (Appendix G). Test item format includes either selected response items or constructed response items.

Application Phase

In this phase diagnostic assessment is formally implemented in the classroom. The instructional strategy used is a teacher-directed, objectives-directed group instruction. On completion of each unit of instruction, a diagnostic test in the form of

a criterion-referenced test is given. The test data are analysed using the item difficulty index p and item discrimination indices B, r_{pb} and ϕ. Students who did not master the unit objectives have a group discussion with the teacher outside the instructional time to rectify their mistakes. Corrective techniques are imparted and problem-solving strategies are stressed.

On completion of all the modules of instruction in the semester, students will sit for a final semester examination. This summative test of 3 hour duration will consist of test items sampled from the domain of instructional objectives for all the modules covered.

To examine whether diagnostic assessment has an effect on test scores in the second semester, the difficulty level and complexity of the first semester tests and second semester tests were rated by two physics instructors in the college. Each judge will rate the tests independently of the other judge, but from the same point of view. 'Independently' means that the judges were expected to exhibit some level of disagreement about the ratings to be awarded, while 'from the same point of view' implies that the judges understood the rating task and apply the rating scale in the same way. The judges were instructed to go through the items in each test and then rate the difficulty and complexity of each test based on the abstractions of the knowledge tested, the intellectual skills demanded based on the hierarchical classification of Bloom's taxonomy, and the test item complexity based on factors within an item which contributes to its difficulty, other than the subject matter and the cognitive skill tested. The ratings of each judge are '1' for 'not difficult', '2' for 'somewhat difficult', 3 for 'quite difficult', and '4' for 'very difficult'. The rating scale is as shown in Appendix I. The means and standard deviations for the ratings by the

judges on the first semester tests and second semester tests are then computed and compared. The purpose is to investigate whether the difficulty levels of the first and second semester tests are similar so that a treatment effect can be suggested.

The mean summative examination scores for the first semester and for the second semester will also be compared to assess the effect of diagnostic assessment on achievement scores in physics.

2. How are students with different achievement scores in physics affected by the mastery versus non-mastery learning strategy employed in diagnostic assessment?

The experimental group of students is classified into 3 levels of achievement in physics based on their first semester final examination scores. The following steps based on the standard deviation method are then used to determine the cut-off scores for classification of students into these levels:

(1) A frequency distribution for the examination scores is built up.

(2) The median and standard deviation of these scores are computed.

(3) The scores that are respectively half a standard deviation higher and lower than the median score represents the upper cut-off score and lower cut-off score for average ability in physics. Scores that are higher than the upper cut-off score are interpreted as high ability while scores below the lower cut-off are low ability scores.

A One-way ANOVA is used to find out whether there are significant differences between the mean scores on the posttest for the three groups of students. If differences exist among the means, a *post hoc* multiple comparison test is used to determine which means differ. In this study, the Tukey's 'honestly significant difference'(HSD) test is used to make all pair-wise comparisons between groups.

For each of the three groups of students, a one-directional matched-groups *t* – test at an α - level of .05 will be used to determine whether there is any significant gain in the mean scores for that group of students as a result of the treatment of diagnostic assessment in the classroom. Each of the *t*- tests will test a null hypothesis that there is no significant difference between the pretest and posttest scores for the respective ability group. The effect size of the treatment for each of the three groups of students is also computed.

The effect of diagnostic assessment on the three pre-classified achievement groups of students is also studied. Single-case designs are used. An interrupted time-series design is used to investigate the influence of diagnostic assessment on these students. The pattern of test scores before and after the introduction of diagnostic assessment serve as repeated measures of achievement scores in physics. A visual inspection of the pattern of test scores before and after the intervention of diagnostic assessment will show the effect of diagnostic assessment on the performance of the three achievement groups of students that are being investigated.

(3) What are the perceptions of students toward the use of diagnostic tests in the classroom?

A Likert-type questionnaire was specifically constructed, piloted and undergone item analysis for the purpose of gathering data on the perceptions of students towards diagnostic tests. This is to ensure that the instrument has high internal consistency. This pilot 5-point rating scale consisting of 59 items is administered to another group of pre-university physics students in the South Australian Matriculation programme. These students are familiar with the diagnostic assessment strategy based on instructional objectives since they were taught by the

4. Is there a significant difference in the perceptions of different physics ability groups toward the classroom diagnostic test?

Areas in diagnostic testing and assessment perceived by the researcher to be effective in identifying students' mastery or non-mastery of instructional objectives may not be true from the student's perspective. Taking the student's point of view may help in detecting flaws and problems in the diagnostic assessment strategy. It is therefore in the student's best interest to be given the opportunity to express his attitudes towards diagnostic tests.

The instrument used to assess students' perceptions towards diagnostic tests is a semantic differential scale that is designed to assess (a) low ability versus high ability students' perceptions of diagnostic tests, and (b) the experimental group's attitudes towards diagnostic tests. This instrument will be administered to the students on completion of the course. The data collected is analysed to answer the specific research questions under consideration.

The following procedure is employed for the purpose of assessing the magnitude of physics achievement group differences in perceptions of the classroom diagnostic test. Specifically, this procedure sets out to test the hypothesis that low achieving students perceive diagnostic test more favorably than high ability students.

The type of instrument used in this study is a Semantic Differential Scale (Appendix B) originally developed by Osgood & Tannenbaum, 1957. It is specifically constructed by the researcher for the purpose of assessing the feelings and perceptions of students towards the classroom diagnostic test. The instrument consists of 18 scales designed to assess five major dimensions. The dimensions and their scales are as below.

(a) Evaluation – good/bad, fair/unfair, pleasant/unpleasant, valuable/worthless, informative/non-informative;

(b) Potency – light/heavy, soft/tough, strong/weak;

(c) Activity – slow/fast, static/dynamic, active/passive;

(d) Comprehensibility – clear/confusing, easy/difficult, simple/complex

(e) Subjective anxiety – threatening/non-threatening, tense/relaxing, fearful/not fearful.

The semantic differential procedure is explained to students. A seven-point rating scale is inserted between bipolar adjectives. This rating scale ranges from 1 to 7. For the five scales in the evaluative dimension, the most negative response (e.g., bad, worthless) were scored 1, the most positive response (e.g., good, valuable)were scored 7. Intermediate strength of responses scored 2, 3, 4, 5 and 6. For the other four dimensions, adjectives in the direction of potent, active, comprehensible and non-anxious rating extremes were scored 7, with the opposite extremes scored 1 respectively. Random polarity of the extreme adjectives is used to reduce the influence of the 'halo effect'.

The low achievers and the high achievers are compared on (a) individual factor scales, (b) each of the five semantic differential dimension scores and (c) their composite mean score on the instrument. The score for each dimension is represented by the mean score of all the factors that represent that dimension. The composite score for each group is the mean score on all the dimensions of the instrument. The inter-correlations between factor scores in the evaluation dimension are computed to support their classification into that dimension. A Cronbach coefficient alpha estimate

of internal consistency was also determined to find out the reliability of the instrument.

The composite mean score for all the students on the instrument is also computed to determine the perceptions of the students towards diagnostic tests in their classroom. According to Campbell and Fiske (1959), if an instrument is a valid measure of attitude toward some object, it should correlate highly with another valid measure of attitude toward the same object; that is, the two measures should exhibit convergent validity.

In this study two measures of the perceptions of students towards diagnostic tests were done using a Likert-type rating scale and a semantic differential scale. A high correlation between these two measures will provide supportive evidence on the concurrent validity of both instruments.

(5) Is there a significant difference in students' perceptions of multiple-choice versus essay type diagnostic tests?

A review of educational evaluation textbooks (Popham, 1981; Sax, 1989; Gronlund & Linn, 1990; Wiersma & Jurs, 1990; Nitko, 1996) shows these authors mainly discuss the various reasons for choosing either essay (constructed-response) or multiple-choice (selected response) item format for classroom tests. These include matters relating to adequate coverage of instructional objectives, ability to measure intended learning outcomes, simplicity of setting test items and ease of scoring, and the reliability of tests based on either format. Little information is available about the student's perceptions towards these two different types of test formats.

Specifically the procedure sets out to test the hypothesis that students perceive multiple-choice diagnostic tests more favorably than essay type tests. The inferential

statistical method used is a one-directional matched groups t – test at an α level of .05 to test the null hypothesis that there is no significant difference between the mean score, μ_1 for multiple-choice test and the mean score, μ_2 for essay type test. The null hypothesis, H_o and alternative hypothesis, H_1 are stated as below.

Null hypothesis

$H_o: \mu_1 = \mu_2$

Alternate hypothesis

$H_1: \mu_1 > \mu_2$

The procedure used involves a 5-point Likert-type rating scale (Appendix C) consisting of two parts which was specifically constructed, piloted and undergone item-analysis for the purpose of gathering data on the perceptions of students towards the two different test formats respectively. Part A of the instrument consists of fourteen 5-point rating scales designed to assess the student's perceptions towards the following fourteen major facets of the multiple-choice diagnostic test. Part B assesses the student's perceptions towards the same fourteen facets of the essay type diagnostic tests.

(a) difficulty (1 = very difficult, 5 = not at all difficult)

(b) complexity (1 = very complex, 5 = not at all complex)

(c) appropriateness (1 = not at all appropriate, 5 = very appropriate)

(d) clarity (1 = not at all clear, 5 = very clear)

(e) value or worth (1 = not at all valuable, 5 = very valuable)

(f) expectancy of success in test (1 = very low, 5 = very high)

(g) effectiveness (1 = not at all effective, 5 = very effective)

(h) degree of interest aroused (1 = not at all interesting, 5 = very interesting)

(i) degree of anxiety (1 = very high anxiety, 5 = very low anxiety)

(j) comfortable feeling (1 = very uncomfortable, 5 = very comfortable)

(k) fairness (1 = very fair, 5 = not at all fair)

(l) fear of failure (1 = very fearful, 5 = not at all fearful)

(m)trickiness (1 = very tricky, 5 = not at all tricky)

(n) degree of motivation aroused (1 = not at all motivating, 5 = very motivating)

The researcher instructs the students to rate the items based on their prior experience with both types of test formats and not to be influenced in their ratings by the specific type of content in these tests.

Scores above 3 for each item suggest favorable perceptions whereas scores below 3 imply unfavorable perceptions. The scores for all the respondents on the two rating scales were summed and the means computed. These mean composite rating scores serve as global measures of perceptions in the statistical analysis.

The mean score for each of the fourteen factors in both scales is analysed to compare the perceptions of students towards that factor.

6. Do different item discrimination indices differ in their effectiveness in discriminating between mastery and non-mastery students on a classroom diagnostic test?

After a diagnostic test has been administered, and the mastery and non-mastery states of the students have been determined based on a cut-off score of 80 percent, a *post hoc* item analysis is performed. For a multiple-choice test, the Brennan discrimination index B (Appendix E), the phi coefficient ϕ (Appendix E) and the point-biserial correlation coefficient r_{pb} (Appendix E) are computed for each item to reflect the degree to which the item discriminates between the group of masters and

the group of non-masters. If the diagnostic test consists of essay type items, then the Brennan discrimination index and the item-total score correlation coefficient (Appendix E) are computed for each item.

For each multiple-choice test, the inter-correlation between the Brennan discrimination indices, the phi coefficients and the point-biserial correlation are computed using a Pearson-r correlation procedure (Appendix E). Similarly, the Pearson-r correlation procedure is used to determine the relationship between the Brennan discrimination indices and item-total correlation coefficients from an essay type test. The range of the Pearson-r's will provide evidence to suggest whether there is a close relationship between the different discriminating indices for the evaluation of teacher-made criterion-referenced tests.

To check whether there is an agreement among different discrimination indices in identifying 'good' and 'poor' items, a discrimination index against difficulty index graph is plotted for each diagnostic test administered to the students. Each item in a test is graphed at the point where its discrimination and difficulty values meet. Lines are drawn on the graph to indicate the optimal item difficulty range for this study (p = .6 to 1.0) with an average difficulty of .8 and the criterion for adequate discrimination (B = .3 , ϕ = .3 or r_{pb} = .3). Test items falling within the section enclosed by these lines represent 'good' items that discriminate adequately between 'masters' and 'non-masters' on the test. The 'poor' items are easy items that do not discriminate. For multiple-choice diagnostic tests, B against p, r_{pb} against p as well as ϕ against p graphs are used to identify 'good' and 'poor' items. For an essay-type test, the B against p and the item-total correlation r against p graphs are used to identify 'good' and 'poor' items. The 'good' and 'poor' items identified

by the different pairs of discrimination indices are correlated using a phi coefficient procedure at a significance level of 5 %. The phi coefficient computed for each pair of indices will provide evidence as to whether there is a significant relationship among the various discrimination indices in identifying both 'good' and 'poor' items in a diagnostic test. This result will also provide an estimate of the concurrent validity possessed by each of the discrimination index.

Instrumentation

The instruments that are developed for this study are:

a) the 5-point Likert-type rating scale (Appendix A) for measuring the perceptions of student toward the use of diagnostic tests,

b) the semantic differential scale (Appendix B) to gather information on the perceptions of different physics ability groups towards diagnostic tests, and

c) the 5-point Likert-type rating scale (Appendix C) to gather data on the perceptions of students towards different test formats of the diagnostic test, and

d) the diagnostic instruments in the form of criterion-referenced tests to measure mastery or non-mastery of instructional objectives for all the units of instruction for the semester.

The Student Questionnaire

The number of items in the pilot questionnaire is too large and may be more than necessary. A *post hoc* item analysis is therefore performed. For each item a correlation is computed between the item and the total score. Only items that correlate strongly with the total score are retained. If the instrument and the item measure the same perceptions, the distribution of scores on the item and the distribution of scores on the instrument should correlate strongly. The researcher decided that a correlation

of magnitude 0.50 or less indicate weak item-total score correlation. The revised Likert-scale is shown in Appendix A. It consists of 36 items of which items 30 and 39 are negative statements.

Semantic Differential Scale

A semantic differential scale is developed to assess the magnitude of physics ability group differences in perceptions of the classroom diagnostic test. For assessing the perceptions of students towards the classroom diagnostic tests, the study chose the following five dimensions: evaluation, potency, activity, comprehensibility and anxiety. The particular bipolar pairs of adjectives are chosen on the basis of the three major underlying dimensions of evaluation, potency and activity, and two lesser dimensions of comprehensibility and anxiety. These dimensions are of interest to the researcher. Since the students' evaluative judgment of diagnostic tests is of interest in this study, his responses to the bipolar pairs in the evaluation dimension such as bad-good, unpleasant-pleasant and worthless-valuable will provide a measure of his perceptions. Similarly bipolar pairs such as light-heavy and slow-fast measure the potency dimension and activity dimensions of the diagnostic tests respectively. The lesser dimensions of comprehensibility and anxiety include bipolar pairs such as easy-hard and tense-relaxed respectively. Student responses will reveal how they view diagnostic tests in terms of the level of difficulty of such tests as well as the degree of their anxious feelings during such tests.

Diagnostic Instrument

Since criterion-referenced tests are used as the instrument of diagnostic assessment of student mastery or non-mastery of instructional objectives, such

instruments need to be developed and validated for classroom use. Two issues are addressed:

(1) Construction of criterion-referenced tests.

(2) Evaluation of criterion-referenced tests.

Construction of criterion-referenced tests involves domain definition, the taxonomy of educational objectives and test specifications.

Domain definition. A criterion-referenced test consists of items that are matched to a set of instructional objectives specified in the domain of instructional objectives for each unit of instruction. The statements of instructional objectives are written in a form and level of specificity that will make them most useful for their intended purpose. Each statement of instructional objectives consists of two parts – an action verb and a content portion.

Taxonomy of educational objectives. The instructional objectives are stated according to the classification of cognitive abilities in Bloom's taxonomy of educational objectives in the cognitive domain. There is a balance between the lower cognitive levels of knowledge, comprehension and the higher levels of application, analysis, synthesis and evaluation.

Test specifications. A key concern in test design is a fair and representative sample of the domain of instructional objectives of the unit of instruction. The purpose of criterion-referenced tests is to evaluate achievement of the objectives for each unit as stated in the domain of instructional objectives. Each objective is considered a domain by itself and a few items are written to measure the attainment of each separate objective.

The following procedures are used to construct criterion-referenced tests in this study.

(1) Definition of content domain: The domain of instructional objectives for each unit of instruction to be measured by the test is defined.

(2) Table of specifications: The content coverage of a criterion-referenced test is outlined and a table of specifications is developed. The table consists of a two-way grid. The instructional objectives of the unit are assigned to rows of the grid while their cognitive levels are assigned to the columns (Appendix G). These cognitive levels are classified based on Bloom's taxonomy in the cognitive domain. The number of test items for each objective are assigned to the cells of the grid.

The table of specifications provides a holistic information on the range of content to be tested and it indicates the relative emphasis to be allocated to each objective. Those objectives that are regarded as essential pre-requisite or prior knowledge for understanding of subsequent instructional units are emphasised by the increased number of items.

(3) Selecting the format of test items

After the test specifications have been developed, the format of test items are decided. In this study, either selected response or constructed response items are used

for each of the tests. In a study comparing selected-and constructed-response versions of analytical reasoning, Bridgeman and Rock (1992) found that the two measures were equivalent. Therefore format does not make a difference when the intent is to measure a common trait. Bennett, Rock and Wang (1990), using achievement tests to compare essay and multiple-choice formats, found a consistently high correlation between the two test formats. Based on their study and an extensive review of literature, they concluded that there is little evidence to show that multiple-choice and constructed-response formats measure substantially different constructs such as factual recall of knowledge versus higher order cognitive processes. Lukhele, Thissen and Wainer (1992), cited in Haladyna (1994), found that for comparable administration times, the essay test yields a lower reliability than the comparable multiple-choice format. This is supported by Wainer and Thissen (1993), who reviewed the history of subjective scoring of essays and found that variation among judges exceeded variability among grades. Hence the multiple-choice format generally produces higher reliability than the essay format. For the one-hour diagnostic test, the essay format cannot provide a very good sample of the content domain, unless that domain is limited. The multiple-choice format permits more items to be administered in a comparable one-hour test. Therefore, the sampling of content is greater than with the use of the essay format. In this study, the multiple-choice format is used when sampling from a large domain of instructional objectives, while an essay-type format is used when the test intends to obtain data on learning difficulties more certainly. This is supported by Martinez (1990) who found that student misunderstanding of the items from an essay test provides information which is not normally available form multiple-choice formats. Therefore some forms of

essay testing provides insights into students' weaknesses in learning. In the context of instructionally based testing, Badger (1990), quoted in Haladyna (1994), argued that even though an essay and its equivalent multiple-choice test are highly correlated, a multiple-choice test is more reliable. On the other hand, the responses to essay items in mathematics and science can reveal the nature of the learning difficulty more specifically. Hence in this study, the diagnostic tests administered to students consist of both formats so as to obtain maximal feedback on student failure in learning the objectives from a unit of instruction. The selected response items used are multiple choice items with four alternatives while the constructed response items consist of questions that require the students to demonstrate a command of essential knowledge and understanding. In scoring essay items, the researcher decided in advance the scoring standards for each part of a given essay question.

Evaluation of the criterion-referenced tests that are used as diagnostic instruments in this study is a process. It involves gathering data about various aspects of the instrument. The researcher focuses on two issues when evaluating such instruments. These are issues relating to the test items and the whole test.

Two types of procedures are used in evaluating the diagnostic tests based on the two issues above. The procedures used are logical review and empirical review. Logical review involves studying the issues before the instrument had been applied while empirical review involves analysing the data which had been collected after application of the instrument.

Test Item Review

A test item must be a quality measure of the domain which is being assessed. Hence, the test item must, firstly, fit the domain definition and, secondly, be free of possible bias which might contaminate the feedback.

Logical Review Procedure

To ascertain the 'fit' and 'quality' of items used in a criterion-referenced test before it is administered to students, two other experienced lecturers in the subject are asked to rate the items of the test on a four-point scale. The questionnaire used in the review of items for a test is shown in Appendix D. The ratings of each judge on each item is dichotomised into weak relevance (ratings of 1 and 2) or strong relevance (ratings of 3 and 4). The number of items that are judged to be strongly relevant by both raters are recorded (Appendix H). A coefficient of content validity equal to the ratio of the number of relevance agreements to the number of items is computed for each diagnostic test. This coefficient of content validity provides evidence on inter-rater reliability.

Empirical Review Procedure

Information about how students performed on a criterion-referenced test can be used to make decisions about the students, the instruction and the item itself. (1). Decision about students: Analysis of the performance of student on each test items would show the pattern of errors for a particular student as well as her or his mastery of instructional objectives that are measured by the items.

Analysis of group performance on each item shows the degree of group mastery of instructional objectives.

(2) Decisions about instruction: Item analysis data is used to make decisions about the pace of instruction. Poor performance on an item indicates that remedial instruction may be needed on the objective measured by the item before moving on to the next unit of instruction.

(3) Decisions about test items: Item analysis data allow us to evaluate the adequacy of each test item. Performance on a test item that is different from others may imply that the item is flawed and need to be reviewed.

The item performance on each diagnostic test is analysed using item difficulty and item discrimination indices. A table showing item performance, including item difficult and the relevant item discrimination indices are constructed after each test, multiple-choice or essay-type format, has been scored.

Item Difficulty

In this study, the difficulty of an item is determined by the proportion of the group of students that was tested who answered the item correctly and is denoted by a difficulty index, p which can range from 0 to +1.

Since p is determined by the responses of test takers to the item, it is a behavioral measure of item difficulty. A difficult item is operationally defined as one that few people answer correctly. Whether an item is a 'good' or 'bad' item depends on the nature of the test and the types of items. Norm-referenced tests are designed to sort test takers into groups relative to each other. Therefore such tests need to include some difficult items, items that challenge the most skillful or able students. Hence a norm-referenced test need to include some items with low difficulty index of .2 or .3.

A norm-referenced test also need to spread out the test scores of the test takers. The variability of test scores is maximised when the values of the difficulty

index average around .5, referred to as the optimal item difficulty for norm-referenced tests. Hence when a test is used to make fine discriminations among test takers, items with difficulty index values near .5 are preferred over items with more extreme values.

In contrast, the criterion-referenced tests used in this study is to determine whether test takers meet some standard of performance. Such diagnostic tests are designed to determine whether test takers have mastered a set of instructional objectives. Test takers are classified into masters or non-masters depending on whether or not they reach the pre-selected criterion. The factors that determine whether a particular test item is difficult or easy depend, among others, on the cognitive skill demanded, the construction of the item itself, and the cognitive ability of the test- taker. The researcher feels that a weak student's response to items of below average difficulty rather than above average difficulty will provide a better indicator of his cognitive ability relative to another better student who responds to the same item. Hence, in this study, the optimal item difficulty range of values is taken to be between .6 to 1.0.

Item Discrimination

The purpose of the diagnostic tests is to discriminate between students who have mastered the domain of instructional objectives that is represented by the test and those who have not achieved mastery. Since the mastery cut-off score for the criterion-referenced tests is 80 %, the researcher wishes to see whether an item discriminates at this cut-off. The item analysis aims to determine how successful did the items on a diagnostic test classify each student as a master or non-master of the domain of instructional objective for a unit of instruction. A set of three item

discrimination indices are computed for each of the items in the multiple-choice diagnostic tests that are administered to students: Brennan discrimination index B, point-biserial correlation coefficient r_{pb} and the phi coefficient ϕ. A set of two discrimination indices B and item-total correlation coefficient is computed for each item in an essay-type test. In this study, the assumption is that instruction is equally effective for all students. Therefore, a positive discrimination index indicates that the item discriminates between 'masters' and 'non-masters'. Adapting from Swezey and Pearlstein (1975), that a phi value of + .30 or higher indicates that the item discriminates well between 'masters' and 'non-masters', the researcher applies the same criterion level for the Brennan discrimination index, the point-biserial correlation coefficient and the item-total correlation. Hence discrimination indices of values ranging from –1.0 and +.30 may imply that the items require some kind of revision or change before they can be included in a criterion-referenced test in future.

Brennan discrimination index. Brennan (1971) introduced an upper-lower discrimination index B (Appendix E). The basic rationale for the index B is that it is the difference between two proportions. A B value of 0.50 is interpreted as 50 percent more students in the upper group got the item correct than in the lower group. Hence a B of 1 is interpreted as all the upper group students got an item correct while none of the lower group students got the item correct. A B value of -1 implies the opposite result. A B of 0 denotes that equal proportions of upper group and lower group students got the item correct.

According to Brennan (1972) the discrimination index B has the advantage of ease of interpretation as compared to correlational indices. In terms of interpretability, a B value of +1 is twice as large as a B value of .50. On the other hand, a

correlation coefficient r *of* $+1$ does not indicate a discrimination that is twice as large as a r of .50.

A negative index denotes that the item is discriminating in the opposite direction from the other items on the test and may be flawed.

Point-biserial correlation. The point-biserial corelation coefficient, r_{pb} is used to correlate dichotomous responses to each item on an objective diagnostic test with the total test scores. A strong significant positive correlation will mean that masters who scored highest on the test tended to get the item right and non-masters who scored low tended to get the item wrong. Therefore analysis of each item will help in identifying items that are doing what the test is supposed to do, to separate the masters from the non-masters.

Phi coefficient. To determine whether there is a correlation between the dichotomous variables of test item scores and mastery/ non-mastery of the test objectives, a *phi coefficient,* ϕ is used. The aim of this procedure is to determine whether or not there is a significant correlation between the state of mastery/non-mastery of the student and his item response.

Test Review

The researcher considered two fundamental issues concerning the whole criterion-referenced tests as opposed to the items which comprised them. These are as follows.

(1) The issue of validity: To what extent does the test measures what it purports to measure?

(2) The issue of reliability: To what extent are the test scores consistent or error-free?

The first issue concerns construct validity of the criterion-referenced tests. It focuses primarily on the test score as a measure of the cognitive aptitude in physics.

The strategy that is adopted for construct validation of the tests follows the suggestions of Campbell and Fiske (1959) and Messick (1989b). Two approaches are used to validate the test – logical review before test administration and empirical review after test administration.

Logical Procedure for Construct Validation

This procedure involves gathering *content-related evidence*. The diagnostic test for each unit is checked to see that it matches important content and learning outcomes to ensure good content coverage. The test items selected are relevant and representative of the domain of instructional objectives for the unit of instruction concerned.

The extent to which the test place emphasis on problem-solving and higher cognitive skills of analysis, synthesis and evaluation are carefully considered to have a good balance between lower and higher order cognitive skills to be measured.

Empirical Procedure for Construct Validation

This procedure involves gathering three types of evidence: internal structure evidence, external structure evidence and reliability evidence.

Internal structure evidence. The diagnostic test for each unit of instruction is checked for reliability. For practical reasons, the Cronbach's alpha procedure for measuring internal consistency is used instead of the test-retest reliability or parallel-form test reliability. In this study the Cronbach's alpha procedure is used for both multiple-choice tests and essay tests. For the multiple-choice tests, the use of Cronbach's alpha provides a more conservative estimate of internal consistency compared to the KR-20 procedure. The students' performance on each diagnostic test

is quantified. Inter-correlations of test scores are computed to see whether there is any pattern of relationships that may emerge.

External structure evidence. The researcher looks for relationships of the diagnostic test scores to the scores on other measures of the same construct that is measured by other tests. The criterion abilities for physics achievement are identified and analysed. In this study the scores on the Physics tests are correlated with the scores on the second semester Further Mathematics examination A strong positive correlation between the two sets of test scores will indicate agreement between the two subject scores in measuring the same construct.

The relationships between the any given measure of a construct with measures of other constructs should be weak. In this study, the scores on the diagnostic tests are correlated with the scores on the second semester Malaysian Studies examination which measures a different construct from that of Physics. A low correlation between the two sets of test scores will suggest that the Malaysian Studies examination is a poor measure of physics ability, and strengthens the construct validity of the physics diagnostics tests.

Each of the diagnostic tests is also correlated with the summative test at the end of the semester to check for predictive validity of each of the diagnostic tests. A strong positive correlation provides evidence on the construct validity of such tests.

Reliability evidence. Reliability evidence is necessary to judge the extent to which measurement errors might interfere with the interpretability of the scores. Sources of errors that affect reliability include variations in the conditions of administration from one testing to the next, differences in scoring or interpreting results, cheating, fluctuations in moods and guessing. The perceptions of the

constructor regarding the match between the test items and the domain of instructional objectives may be inconsistent. To overcome this possible threat to content-related validity of the tests , two experts in the subject are asked to rate the appropriateness of the items. A high correlation between the two raters denotes the high inter-rater reliability. To demonstrate the reliability of a test, a reliability coefficient is computed. The internal consistency reliability is computed for each criterion-referenced test in this study. Test-retest and parallel form test reliabilities are impractical for use in the classroom. The Cronbach's alpha procedure is used for the objective tests that are scored dichotomously (0 or 1), one point for each correct answer and no points for an incorrect answer. The Cronbach alpha procedure is also used for estimating the reliability of essay test scores.

Test Reliability

The second issue that need to be addressed in reviewing a criterion-referenced test is the issue of reliability. Reliability of a test refers to the consistency of the test in measuring whatever it is measuring. According to Ebel and Frisbie (1991), the reliability of a set of test scores from a group of examinees is measured by the correlation between that set of scores and another set of scores on an equivalent test obtained independently from the members of the same group. Wiersma and Jurs (1990) discusses three types of reliability: stability reliability, equivalence reliability and internal consistency reliability.

Stability reliability. Stability reliability refers to consistency of measurement across time. The procedure for estimating stability reliability is the test-retest. The reliability coefficient is the correlation coefficient between the scores of the two test administrations.

Equivalence reliability. Equivalence reliability refers to consistency of measurement across two parallel forms of a test. The reliability coefficient is based on the correlation between two parallel forms of a test administered at the same time.

Internal consistency reliability. The administration of test-retest and parallel forms of a test is impractical for classroom tests. A procedure based on a single test administration is used in this study. The items and the scores on a test are split into two halves, and a measure of internal consistency reliability is computed by correlating the two halves. Since the split-half coefficient is an underestimate of the actual reliability of a test, the Spearman-Brown prophesy formula (Appendix E) is used to estimate the reliability of the complete test. For most tests, there is a large number of possible splits. The KR-20 formula (Appendix E) gives an estimate of internal consistency reliability r_{20} equal to the mean of all possible split-half coefficients. The KR-20 formula is used to review the reliability of tests involving dichotomously scores test items. For the purpose of reviewing the reliability of essay type test, the Cronbach alpha procedure (Appendix E) is used. The alpha coefficient does not require the right-wrong dichotomous scoring of individual items. The Cronbach alpha procedure will generate a more conservative estimate of a test's internal consistency compared to the KR-20 procedure because it compares performance on each test item to performance on all other items. In this study, the Cronbach's alpha procedure is used to estimate the reliability of the pretest and posttest although they are summative tests that cover a heterogeneous domain of instructional objectives. This may yield a lower estimate of internal consistency. The Cronbach's alpha procedure is also used for the multiple-choice and essay-format

diagnostic tests. In such tests, all the test items are homogeneous since they are drawn from the same domain of content knowledge.

Standard Error of Measurement

For each of the diagnostic test scores, a standard error of measurement (SEM) is computed. The SEM (Appendix E) is an index of the average amount of error in test scores. In practice, SEM uses the reliability coefficient to determine the average number of points by which test scores differ from true scores. In a diagnostic test, a distribution of test scores is obtained. In this study, the SEM is interpreted in the context of the test score scale. For example, a SEM of 5 points on a test with a maximum score of 100 points is not significant as compared to a typical diagnostic multiple-choice diagnostic test in this study in which the top score is about 20 points.